STAT 401 Chapter 5.8–5.11, 5.13

Zepu Zhang
November 4, 2010

1 Random vectors/matrices; covariance matrix

In our linear model, if we arrange the Y's corresponding to a dataset in a vector, we can write

$$\vec{Y} = \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix}$$

Similarly, we may put all the random errors in a vector:

$$\vec{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

Note, since the elements of $\vec{Y}$ and $\vec{\epsilon}$ are random variables, $\vec{Y}$ and $\vec{\epsilon}$ are random vectors.

Similarly, there are random matrices.

Now back out of the linear regression context. Use $\vec{X}$ to denote a general random vector. We can write $\vec{X} = [X_1, X_2, \ldots, X_n]^T$, because a vector is generally understood as a column vector (or matrix).

The mean $E(\vec{X})$ is the vector composed of the mean of each element of $\vec{X}$, that is

$$E(\vec{X}) = \begin{bmatrix} E(X_1) \\ \vdots \\ E(X_n) \end{bmatrix}$$

Of course, we can talk about the element-wise variances. But, we'll go more general and talk about the covariance between every pair of elements of $\vec{X}$. Define the covariance matrix of the random vector (or multivariate random variable) $\vec{X}$ as

$$\operatorname{cov}(\vec{X}) = \begin{bmatrix} \operatorname{cov}(X_1, X_1) & \ldots & \operatorname{cov}(X_1, X_n) \\ \vdots & \ddots & \vdots \\ \operatorname{cov}(X_n, X_1) & \ldots & \operatorname{cov}(X_n, X_n) \end{bmatrix}$$

Therefore, the $(i,j)-$th element of the covariance matrix is $\text{cov}(X_i, X_j)$ whereas the (i,i)-th is $\text{cov}(X_i, X_i)$, that is, $\text{var}(X_i)$.

(There is no standard symbol for this matrix. Just define it in your context. $\boldsymbol{\Sigma}$ is a common choice. The symbol used in KNNL, $\boldsymbol{\sigma}^2$, is awkward. Always use an upper-case letter for matrix, unless the matrix is a row or column vector.)

From $\text{cov}(X_i, X_j) = \text{cov}(X_j, X_i)$ we see $\underline{\text{cov}(\vec{X}) \text{ is symmetric}}$.

We also know $\det(\text{cov}(\vec{X})) > 0$. Therefore a $\underline{\text{covariance matrix is always invertible}}$. In statistics, this determinant is often written as $|\text{cov}(\vec{X})|$ (or $|\boldsymbol{\Sigma}|$, if $\boldsymbol{\Sigma}$ denotes the covariance matrix).

By the definition of covariance (between two univariate random variables), we see

$$\text{cov}(\vec{X}) = E\Big[\big(\vec{X} - E(\vec{X})\big)\big(\vec{X} - E(\vec{X})\big)^T\Big]$$

that is, it is the expectation of a random matrix whose elements are "centered cross-products".

Linear combinations of random variables

Let $\vec{X}$ be a n-random vector and $\vec{Y} = \boldsymbol{A}\vec{X}$ where $\boldsymbol{A}$ is $m \times n$. Then $\vec{Y}$ is a m-random vector (m can be 1).

Exercise A matrix is tool for describing $\underline{\text{linear}}$ combinations. Explain how the definition of $\vec{Y}$ expresses each element of $\vec{Y}$ as a linear combination of the elements of $\vec{X}$.

$\underline{\text{The expectation of } \vec{Y} \text{ is}}$

$$E(\vec{Y}) = E\big(\boldsymbol{A}\vec{X}\big) = \boldsymbol{A}\,E(\vec{X})$$

$\underline{\text{The covariance of } \vec{Y} \text{ is}}$

$$\text{cov}(\vec{Y}) = \text{cov}\big(\boldsymbol{A}\vec{X}\big) = \boldsymbol{A}\,\text{cov}(\vec{X})\boldsymbol{A}^T$$

Exercise Verify the sizes of $\text{cov}(\vec{Y})$ and $\boldsymbol{A}\,\text{cov}(\vec{X})\boldsymbol{A}^T$ are identical.

$\underline{\textbf{Important result}}$: If $\vec{X}$ is a normal random vector (i.e., the elements of $\vec{X}$ have a joint multivariate normal distribution), then $\vec{Y}$ is also a normal random vector. Recall a normal distribution is completely describe by the mean and variance (or covariance matrix, if multivariate).

$$\vec{X} \sim N(\vec{\mu}_X, \boldsymbol{\Sigma}_X) \quad \Rightarrow \quad \vec{Y} = \boldsymbol{A}\vec{X} \sim N\big(\boldsymbol{A}\vec{\mu}_X, \boldsymbol{A}\boldsymbol{\Sigma}_X\boldsymbol{A}^T\big)$$

Example Page 194.

Example Page 196.

2 SLR Model formulation in Matrix

$$Y = \beta_0 + \beta_1 X + \epsilon$$

For this model there is no need to use matrix. It's clear, simple, and complete.

Matrix notation kicks in when we want to (concisely) present the model behind a data set, (X_i, Y_i), $i = 1, \cdots, n$, because now there is one model for each Y_i.

Define the notation $\vec{Y} = [Y_1, \ldots, Y_n]^T$ (random response vector),

$$\boldsymbol{X} = \begin{bmatrix} 1 & X_1 \\ \vdots & \vdots \\ 1 & X_n \end{bmatrix}$$

("design matrix"), $\vec{\beta} = [\beta_0, \beta_1]^T$ (coefficient vector), and $\vec{\epsilon} = [\epsilon_1, \ldots, \epsilon_n]^T$ (rand errors). The SLR model for the dataset is

$$\vec{Y} = \boldsymbol{X}\vec{\beta} + \vec{\epsilon}, \quad \text{with assumption} \quad \vec{\epsilon} \sim N(\vec{0}, \sigma^2 \boldsymbol{I}).$$

The assumptions on $\vec{\epsilon}$ (zero mean, constant variance, independent normal) are concisely and fully represented by $\vec{\epsilon} \sim N(\boldsymbol{0}, \sigma^2 \boldsymbol{I})$. Note σ^2 is a scalar (single number, not bold).

Note On notation:

- Usually we use capital letters for matrices. Preferably bold. (You will see books that use non-bold capital letters for matrices.)
- We often use bold or arrowed lower-case letters for vectors. (However, since vectors are matrices, in principal we could use the matrix notation for vectors.) I guess the arrowed form is a legacy from the typewriter or handwriting days.
- We use upper-case letters for random variables, hence the random vector $\vec{Y}$. (But it's not bold. For vector, we use either bold or arrow; there is no need for both decorations.)
- The design matrix $\boldsymbol{X}$ is not random; it's all fixed numbers.

- Use either ^{T} or $'$ for matrix transpose.
- A little ambiguity in the notation: we use $\vec{Y}$ for the random vector of Y's (behind the observed Y values), and also for the actual observations.

3 LS estimators

The LS condition leads, by setting to zero the derivative of $\sum e_i^2$ w.r.t. every parameter, to a group of "normal equations" (equations (1.9) on page 17). Write this system of equation in matrix form, we get

$$\boldsymbol{X}'\boldsymbol{X}\vec{\hat{\beta}} = \boldsymbol{X}'\vec{Y}. \tag{1}$$

See the section "**Normal Equations**" on page 199.

Exercise Check this matrix equation does mean the same thing as the equation system (1.9).

Note: any linear equation system is just one line in matrix notation. Equation (1) has three components:

- coefficient matrix: $\boldsymbol{X}'\boldsymbol{X}$
- vector of unknowns: $\vec{\hat{\beta}}$
- vector of constants: $\boldsymbol{X}'\vec{Y}$

The solution to (1) is

$$\vec{\hat{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\vec{Y} \tag{2}$$

assuming the inverse exists.

Q Does it happen often that $\boldsymbol{X}'\boldsymbol{X}$ is not invertible?

Now it's not obvious how we can verify this solution is the same as the separate expressions of $\hat{\beta}_0 = ...$ and $\hat{\beta}_1 = ...$ we have seen. However, this does not need to be verified (although you're welcome to try). As long as the matrix notation for the linear system is correct, the solution must be correct. (In fact, the matrix notation of linear system and its solution is more standard than the previous way of doing things in separate lines.)

Here's a better (and more direct) way to derive (2). The LS criterion is to minimize the total squared errors:

$$Q = \sum e_i^2 = \vec{e}^T\vec{e} = \left(\vec{Y} - \boldsymbol{X}\vec{\beta}\right)^T\left(\vec{Y} - \boldsymbol{X}\vec{\beta}\right) \tag{3}$$

Note Q is just a number, although it is expressed through matrices here. The solution should be obtained by setting to zero the derivative of Q wrt each element of $\vec{\beta}$. If we look at a specific element β_i, then the derivative is

$$\begin{aligned}\frac{\partial Q}{\partial \beta_i} &= \frac{\partial\left(\vec{Y}-\boldsymbol{X}\vec{\beta}\right)^T}{\partial \beta_i}\left(\vec{Y}-\boldsymbol{X}\vec{\beta}\right) + \left(\vec{Y}-\boldsymbol{X}\vec{\beta}\right)^T \frac{\partial\left(\vec{Y}-\boldsymbol{X}\vec{\beta}\right)}{\partial \beta_i} \\ &= \left(-\boldsymbol{X}_{[,i]}\right)^T\left(\vec{Y}-\boldsymbol{X}\vec{\beta}\right) + \left(\vec{Y}-\boldsymbol{X}\vec{\beta}\right)^T\left(-\boldsymbol{X}_{[,i]}\right) \\ &= -\boldsymbol{X}_{[,i]}^T\vec{Y} + \boldsymbol{X}_{[,i]}^T\boldsymbol{X}\vec{\beta} - \vec{Y}^T\boldsymbol{X}_{[,i]} + \vec{\beta}^T\boldsymbol{X}^T\boldsymbol{X}_{[,i]} \\ &= 2\boldsymbol{X}_{[,i]}^T\boldsymbol{X}\vec{\beta} - 2\boldsymbol{X}_{[,i]}^T\vec{Y}\end{aligned} \tag{4}$$

where $\boldsymbol{X}_{[,i]}$ represents the i-th column vector of $\boldsymbol{X}$. Notice that all the four terms on the second-to-last line are scalars. Further notice that the transpose of a scalar is itself, leading to the combinations in the last line.

Denote the "derivative vector" $\left[\frac{\partial Q}{\partial \beta_1}, \ldots\right]^T$ by $\frac{\partial Q}{\partial \vec{\beta}}$, then we see

$$\frac{\partial Q}{\partial \vec{\beta}} = 2\boldsymbol{X}^T\boldsymbol{X}\vec{\beta} - 2\boldsymbol{X}^T\vec{Y} \tag{5}$$

Set this derivative vector to $\vec{0}$, and suppose $\hat{\vec{\beta}}$ is the solution, then

$$\boldsymbol{X}'\boldsymbol{X}\hat{\vec{\beta}} = \boldsymbol{X}'\vec{Y} \tag{6}$$

leading to

$$\hat{\vec{\beta}} = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\vec{Y}$$

Remarks

1. If you are careful and patient, you can verify (4). It's not that special; both Q and β_i are just scalars.

2. If you are comfortable with matrices, however, you would skip (4) and go from (3) to (5). Not totally straightforward, but there are formulas for matrix derivatives to look up. (For example, `http://matrixcookbook.com`)

3. Notational detail: we changed $\vec{\beta}$ to $\hat{\vec{\beta}}$ while going from (5) to (6) because it is the at the estimators $\hat{\vec{\beta}}$ (instead of the unknown, true $\vec{\beta}$) that the derivatives are zero.

4. The point to make is: we get the equation (6) not by translating a system of equations like what we derived in Chapter 1, but from the matrix version of the model directly.

5. The advantage? If we have multiple predictors, then the matrix $\boldsymbol{X}$ will have more columns and the coefficient vector $\vec{\beta}$ will have more entries. However, the LS criterion (3) stays the

same, and we have not restricted the size of either $\boldsymbol{X}$ or $\vec{\beta}$ in the arguments leading to (6), therefore the solution stays the same.—Yah! That's multiple regression and we have solved it, in a completely general form.

Tip How can I memorize this solution? Here's how I do it:

Think of $\boldsymbol{X}\vec{\hat{\beta}} = \vec{Y}$. We do not have such an equality (because there won't be a $\vec{\hat{\beta}}$ that makes this hold unless all points (x_i, y_i) fall exactly on a straight line), but it is roughly what we mean by the model. Starting with this, we want to solve it but it can't be done, because the coefficient matrix $\boldsymbol{X}$ is not square. Then we make it square by left-multiplying $\boldsymbol{X}'$ on both sides, getting $\boldsymbol{X}'\boldsymbol{X}\vec{\hat{\beta}} = \boldsymbol{X}'\vec{Y}$. The solution is then straightforward (and standard).

Example Page 200.

4 Fitted values

$$\vec{\hat{Y}} = \boldsymbol{X}\vec{\hat{\beta}} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\vec{Y} \tag{7}$$

Let $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$, then

$$\vec{\hat{Y}} = \boldsymbol{H}\vec{Y} \tag{8}$$

Clearly from here, the fitted values are linear functions of the observed values. The link between the two is the matrix $\boldsymbol{H}$. $\boldsymbol{H}$ is known as the "hat" matrix, and is useful in some discussions.

You should remember the definition of $\boldsymbol{H}$, and use it to simplify things. For example, don't memorize (7); instead, memorize (8) and know what $\boldsymbol{H}$ is.

Exercise Show that $\boldsymbol{H}$ is symmetric.

Exercise Show that $\boldsymbol{H}\boldsymbol{H} = \boldsymbol{H}$.

Example Page 202.

Since $\vec{\hat{Y}}$ is a linear function of $\vec{Y}$, the former has a normal sampling distribution with mean

$$E(\vec{\hat{Y}}) = \boldsymbol{H}E(\vec{Y}) = \boldsymbol{H}\boldsymbol{X}E(\vec{\hat{\beta}}) = \boldsymbol{X}\vec{\beta}$$

and covariance matrix

$$\operatorname{cov}(\vec{\hat{Y}}) = \boldsymbol{H}\operatorname{cov}(\vec{Y})\boldsymbol{H}' = \boldsymbol{H}\,\sigma^2\boldsymbol{I}\,\boldsymbol{H}' = \sigma^2\boldsymbol{H}$$

If we want the fitted value at a predictor x_* (not necessarily in the observations), then let $\boldsymbol{X}_* = [1, x_*]$ and

$$\hat{Y}_* = \boldsymbol{X}_* \vec{\hat{\beta}}$$

This is again normal with mean and variance

$$E(\hat{Y}_*) = \boldsymbol{X}_* E(\vec{\hat{\beta}}) = \boldsymbol{X}_* \vec{\beta}, \quad \operatorname{var}(\hat{Y}_*) = \boldsymbol{X}_* \operatorname{cov}(\vec{\hat{\beta}}) \boldsymbol{X}_*'$$

We'll derive $\operatorname{cov}(\vec{\hat{\beta}})$ in a moment.

5 Residuals

$$\vec{e} = \vec{Y} - \vec{\hat{Y}} = (\boldsymbol{I} - \boldsymbol{H})\vec{Y}$$

This shows that the residuals are also linear functions of the observations (the vector $\vec{Y}$).

Exercise Express SSE using $\boldsymbol{X}$ and $\vec{Y}$.

$$\begin{aligned}
\text{SSE} = \sum_i e_i^2 &= \vec{e}'\vec{e} \\
&= \left[(\boldsymbol{I} - \boldsymbol{H})\vec{Y}\right]'(\boldsymbol{I} - \boldsymbol{H})\vec{Y} \\
&= \vec{Y}'(\boldsymbol{I} - \boldsymbol{H})'(\boldsymbol{I} - \boldsymbol{H})\vec{Y} \\
&= \vec{Y}'(\boldsymbol{I} - \boldsymbol{H} - \boldsymbol{H}' + \boldsymbol{H}'\boldsymbol{H})\vec{Y} \\
&= \vec{Y}'(\boldsymbol{I} - \boldsymbol{H})\vec{Y}
\end{aligned}$$

Remember, $\hat{\sigma^2} = \text{MSE} = \text{SSE}/(n-2)$.

6 Inferences in matrix notation

Look at the solution

$$\vec{\hat{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\vec{Y}.$$

This is a linear function of $\vec{Y}$, which is a normal random vector, therefore $\vec{\hat{\beta}}$ is normal, with mean

$$E(\vec{\hat{\beta}}) = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}' E(\vec{Y}) = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X}\vec{\beta} = \vec{\beta}$$

and covariance matrix

$$\begin{aligned}
\operatorname{cov}(\vec{\hat{\beta}}) &= \left[(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\right] \operatorname{cov}(\vec{Y}) \left[(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\right]' \\
&= \left[(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\right](\sigma^2\boldsymbol{I})\left[(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\right]' \\
&= \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1} \\
&= \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}.
\end{aligned}$$

Remarks 1. The manipulations of $\boldsymbol{X}$ in the above should become very familiar to you. In particular, you should train yourself to see readily that
(1) $\boldsymbol{X}'\boldsymbol{X}$ is symmetric;
(2) $\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}$ is symmetric;
(3) $\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{X} = \boldsymbol{I}$.

2. The covariance matrix of $\vec{\hat{\beta}}$ above gives not only the variances of $\hat{\beta}_0$ and $\hat{\beta}_1$, but their covariance $\text{cov}(\hat{\beta}_0, \hat{\beta}_1)$. In Chapter 2 we thought the dependence between $\hat{\beta}_0$ and $\hat{\beta}_1$ was hard to describe but we walked around it. Now their covariance is simply the off-diagonal element of the matrix $\text{cov}(\vec{\hat{\beta}})$. (And this idea generalizes however many elements β contains.)

Example Page 207.

In a totally analogous fashion, we can make inferences about the expected value of Y at given X, say x_*. Let $\boldsymbol{X}_* = [1, x_*]$, then

$$\hat{Y}_* = \boldsymbol{X}_* \vec{\hat{\beta}}$$

This is again normal with mean and variance

$$E(\hat{Y}_*) = \boldsymbol{X}_* E(\vec{\hat{\beta}}) = \boldsymbol{X}_* \vec{\beta}, \quad \text{var}(\hat{Y}_*) = \boldsymbol{X}_* \,\text{cov}(\vec{\hat{\beta}}) \boldsymbol{X}_*' = \sigma^2 \boldsymbol{X}_* (\boldsymbol{X}'\boldsymbol{X})^{-1} \boldsymbol{X}_*'$$

Exercise We learned in Chapter 2 that $\text{var}(\hat{Y}_*) = \sigma^2 \left(\frac{1}{n} + \frac{(x_* - \overline{x})^2}{S_{xx}} \right)$. Check this is identical to the above. To check this we need to use the formula for the inverse of a 2×2 matrix, $\boldsymbol{X}'\boldsymbol{X}$ here.

Example Page 208.